CHAOS: User-driven Development of a Metadata Scheme for Radio Broadcast Archives

Lykke, Marianne; Bogers, Toine; Larsen, Birger; Lund, Haakon

Published in:
iConference 2013 Proceedings

DOI (link to publication from Publisher):
10.9776/13510

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

CHAOS: User-driven Development of a Metadata Scheme for Radio Broadcast Archives

Haakon Lund
Royal School of Library and Information Science, Copenhagen
hl@iva.dk

Toine Bogers
Royal School of Library and Information Science, Copenhagen
tb@iva.dk

Birger Larsen
Royal School of Library and Information Science, Copenhagen
bl@iva.dk

Marianne Lykke
University of Aalborg
Department of Communication and Psychology
mlykke@hum.aau.dk

Abstract

CHAOS[1] (Cultural Heritage Archive Open System) is a digital platform for Danish radio broadcasts. Radio broadcasts are an important and vibrant part of our cultural heritage, but providing efficient and effective access to such archives is challenging for lack of a solid digital infrastructure. The Danish LARM project aims to meet this challenge by making one million hours of radio programs available to humanities researchers through the digital platform CHAOS. CHAOS is being built in close cooperation with the researchers involved in LARM. In this paper, we present the user-driven development of the multi-tiered metadata scheme used in CHAOS.

Keywords: metadata, cultural heritage, user-driven development, radio, digital humanities

Introduction

The past two decades have seen mass investments in large-scale digitization projects aimed at digitally preserving our cultural heritage. In more recent years, the focus of such efforts has turned to the question of how to provide efficient and effective access to these digitized collections for both academia and the general public. One such project is the Danish LARM project[2], which is a joint initiative between the Danish national broadcasting corporation (DR)[3], the State and University Library (SB)[4] hosting the Danish Media Archive, and a consortium of Danish university humanities departments.

The goal of LARM is to unlock the true potential of Danish digitized radio broadcast archives by providing access to over a million hours of radio to humanities researchers through a dedicated digital infrastructure called CHAOS. CHAOS offers streaming access to all the material broadcast in the period 1985-2005 and selected broadcasts dating all the way back to 1925. Examples of the diverse range of use cases the project caters to, are investigating the changes in the Danish language as used on the radio, and analyzing the changing rhythm and pace of radio broadcasts throughout the years.

[1] See http://www.larm-archive.org/about-larm/about-chaos/ for more information
[2] See http://www.larm-archive.org/about-larm for more information.
[3] http://www.dr.dk
[4] http://en.statsbiblioteket.dk/about-the-library

Acknowledgements: The LARM-project is supported by The National Programme for Research Infrastructure (Grant number 09-067292). We thank Assistant Professor Mette Skov, Royal School of Library and Information Science, for fruitful discussion in developing the meta data scheme.

Lund, H, Bogers, T, Larsen, B & Lykke, M. (2013). CHAOS: User-driven Development of a Metadata Scheme for Radio Broadcast Archive. *iConference 2013 Proceedings* (pp. 990-994). doi:10.9776/13510

An integral part of enabling efficient and effective access to these radio broadcasts is developing a metadata scheme for handling the archival data related to these broadcasts that can both capture the original metadata and handle the diverse user needs of the participating researchers. In this paper we present work on the user-driven development of a metadata scheme that can satisfy the needs of all the LARM partners and unlock the true potential of Danish radio archives.

User Requirements

Development of the CHAOS platform and infrastructure is characterized by a bottom-up approach where development takes place in close cooperation with the researchers taking part in the LARM project. This user-driven approach has also been an essential part of developing the metadata scheme used in LARM, which has the effective and efficient retrieval of broadcasts as its main objective. DR's own broadcast archives contain archival metadata assigned at the level of individual broadcasts. Much of this *core metadata* has been transferred to the CHAOS archives, although some metadata fields could not be transferred due to copyright and privacy concerns. Therefore core metadata in CHAOS is sparsely populated, with a positive bias towards later years, and does not allow for the effective retrieval of radio content.

To optimally support the humanities researchers in their work, we performed a *user requirements analysis* to determine how this metadata scheme should be extended. We performed this requirements analysis through a series of workshops with the participating researchers and through an online survey. The main purpose of these workshops was to gauge the researchers' expectations with regard to descriptive archival metadata, and to identify which types of additional annotations researchers deemed necessary to support them in their daily research practice. We found that researchers expressed a desire for not only metadata that supports the effective retrieval of radio broadcasts, but also for adding research-specific annotations at both the broadcast level as well as at segments of broadcasts. A main outcome of the requirements analysis is that the needs of humanities researchers are so diverse that it is unlikely that a single unified metadata list will suit all. As a consequence, our proposed metadata scheme includes project-specific metadata, allowing individual researchers and project to adapt metadata lists to their needs (Skov & Lykke, 2012).

Conceptual Metadata Scheme

Our main objective in creating the CHAOS metadata schema was to develop a metadata scheme that was easy to work with, easily extensible, and would provide for flexible data exchange. For this reason, the CHAOS metadata scheme was built on top of the Dublin Core Metadata Element Set (Dublin Core Metadata Initiative, 2012) as implemented by the European Broadcasting Union (EBU, 2011). Using Dublin Core as our foundation also simplifies data exchange with other cultural heritage repositories, such as Europeana (Europeana, 2012)

Our user requirements analysis revealed the following three levels of metadata relevant to LARM researchers: (1) core archival metadata, (2) LARM metadata, and (3) project-specific metadata. For each of the three levels of descriptive metadata a number of administrative metadata were identified. In addition, researchers expressed a desire for functionality in CHAOS that could support them in their annotation. Figure 1 illustrates the resulting metadata architecture developed for CHAOS based on these requirements. We describe each of these metadata levels and the annotation support systems in more detail in the rest of this section.

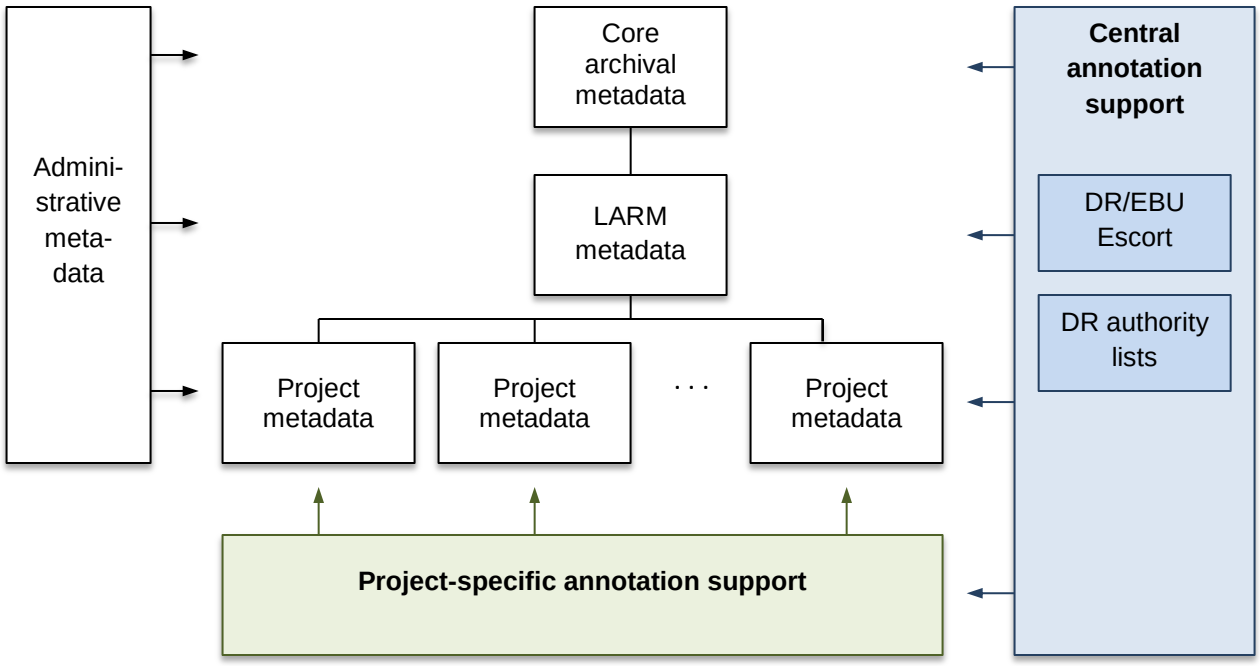

Figure 1. Proposed metadata architecture in CHAOS. The architecture contains three different levels of metadata and provides for both central and project-specific annotation support.

Core Archival Metadata

Core archival metadata covers metadata inherited from the original data source—usually DR or the State and University Library. This core metadata is immutably tied to each broadcast and is regarded as historical data, with all the possible flaws this might contain. Core metadata is assigned at the level of individual broadcasts and is intrinsic to the broadcast. The amount of data is limited to technical information for the most part, with a few descriptive additions, e.g., program title, producer, and a short narrative. Below is an example of some core archival metadata elements:

Metadata element	Description	Example
Radio channel	Broadcast channel	P2
Program title	Original title	Til Italien!
Program start (time)	Date and time for beginning of program	18. mar. 2006 kl 19:00:00
Program end (time)	Date and time for end of program	18. mar. 2006 kl 19:30:00
Narrative	Description of content	Mendelssohn: Symfoni nr 4. Den Italienske. Stuttgarts Radiosymfoniorkester. Dirigent: Roger Norrington
ID of origin	Unique Production ID	
Creator and role	Producer, etc.	Hans Hansen (Producer)

LARM Metadata

LARM metadata is descriptive metadata with the aim of enriching the sparse core metadata with more detailed information about content and participants of a broadcast. LARM metadata is also assigned at the level of individual broadcasts and is editable by all LARM members. Typically, the first researcher

to use the radio program for research purposes adds this metadata. Below is an example of some LARM metadata elements:

Metadata element	Description	Example
Program title	If title at the archival level is absent or incomplete.	
Person - participant	From help system (controlled)	Roger Norrington
Person - subject	From help system (controlled)	Felix Mendelsohn Bartholdy
Genre	From help system (controlled)	Koncertoptagelse
Related objects	Webpage, podcast, photo, etc.	URL to ressource
Subject	From help system (controlled)	Klassisk musik
Tag	User defined keyword (uncontrolled)	
Annotation	Annotation related to entire show	

Project-specific Metadata

Project-specific metadata is of a more analytical nature and is associated with one or more individual research projects within the LARM initiative. Project-specific metadata can describe an entire show or parts of a show. Project metadata is assigned by the researchers working on a LARM research project, and is owned by that research project. Metadata at the project level is designed to be open and flexible, but the following metadata elements are suggested:

Metadata element	Description	Example
Title	Object title	Introduktion til koncerten
Person - participant	Person participating in show	Magnus Møller
Person - subject	Person as subject for the show or part of show	
Genre	Project defined genre (from help system)	Speak
Related objects	Webpage, podcast, photo etc.	
Subject	Project defined subject (from help system)	
Tag	Project defined subject (uncontrolled)	
Object start	Time for start of part of program	19:00:00
Object end	Time for end of part of program	19:05:30
Annotation	Project annotation to entire show or part of show	

Controlled keywords and genre information could originate from an existing support system within CHAOS or each project could define their own authority lists.

Administrative Metadata

An important part of a working metadata system is administrative metadata. In CHAOS, administrative metadata is related to each of the three metadata levels and provides information about the metadata record, i.e., when and by whom the record was created. This data can be generated automatically, e.g., from user login. Below are some examples of administrative metadata related to the archival level:

Metadata element	Description	Example
ID	Unique ID in CHAOS	
Source system	Originating source system	DR
Original ID	Unique record ID in source system	7776671
Date	Date and time for injection in CHAOS	

Administrative metadata at LARM and project-specific levels should also include information about when and by whom records have been edited. Here, administrative metadata does not only have to be related to the record as a whole, but can also apply to actions taken on individual metadata elements in the metadata records, i.e., who changed the title information and when.

Annotation Support

Our proposed metadata architecture (see Figure 1) provides for different help systems to support users in their data entry and annotation process. Such support systems could take the form of standardized vocabularies or taxonomies, authority lists of person names, etc. Their main purpose is to secure uniform entry points to help users in the retrieval of radio material.

We propose two types of support systems: a centralized support system as well as project-specific support. The centralized support system allows for the importing of existing taxonomies and authority lists, such as EBU Escort (EBU, 2007) or DR index terms. In addition, each project should be able to add its own project-specific support system. Every support system should be documented by means of a coding manual. Documentation about existing standardized system, such as taxonomies, should be in the form of a reference to an existing formal documentation or standard. Project-specific support systems should provide the required documentation themselves, including information about their intended use, coding, etc. We propose that a specific interface be implemented in CHAOS for the creation and editing of project-specific support systems and the automatic importing of existing classification schemes. The *"EBU Reference Data & Classification Schemes"* (EBU, 2011) could be used as reference scheme for this.

Conclusion & Future Work

The proposed CHAOS metadata scheme has been implemented in the latest release of the CHAOS infrastructure. We aim to conduct a thorough evaluation of the metadata scheme to qualify its relevance. This will include both an evaluation of the suggested metadata elements, but also of the CHAOS user interface.

References

Skov, M. & Lykke, M. (2012). Unlocking radio broadcast: user needs and sound retrieval. *Proceedings of the 4th Information Interaction in Context Symposium*. New York, NY: ACM

Dublin Core Metadata Initiative (2012) *Dublin Core Metadata Element Set, Version 1.1*. Retrieved from http://dublincore.org/documents/dces/

EBU (2011) *EBU core metadata set*. (EBU-TECH 3293). Retrieved from http://tech.ebu.ch/docs/tech/tech3293v1_3.pdf

Europeana (2012). Retrieved from http://www.europeana.eu/portal/, last visited September 24, 2012

EBU (2007) *ESCORT 2007. EBU system of classification of radio and television programmes.* (EBU-TECH 3322). Retrieved from http://tech.ebu.ch/docs/tech/tech3322.pdf

EBU (2011) *EBU reference data & classification scheme.* (EBU-TECH 3336). Retrieved from http://tech.ebu.ch/docs/tech/tech3336v1_1.pdf